Corpus Linguistics and English Across 'The Three Circles'

Corpus Linguistics and English Across 'The Three Circles':

A Student's Guide

By

Rita Calabrese

Cambridge
Scholars
Publishing

Corpus Linguistics and English Across ‘The Three Circles’:
A Student’s Guide

By Rita Calabrese

This book first published 2024

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-7498-9
ISBN (13): 978-1-5275-7498-4

CONTENTS

INTRODUCTION

This volume provides a survey of issues and studies on ‘applied’ corpus linguistics across two crucial decades which have marked enormous advancement in the field of corpora studies. An early Italian edition of the volume appeared in 2004 when corpus linguistics studies were just emerging in Italy as a novel, challenging, intriguing framework of new linguistic theories and alternative methods of language investigation. At present, corpus linguistics and its applications form a well-established field of research which deserves special attention by both EFL students and practitioners actively engaged in the study of the English language. The first edition reported on a project which was carried out at the Department of Foreign Languages and Literary Studies of the University of Salerno in the academic years 2002/2004. The main purpose of the project was the creation and related analysis of a corpus of English tests written by Italian students of English as a Foreign Language (EFL). The essential aim of the research was to define a taxonomy of the most recurrent errors along with the identification of the most typical features of their interlanguage. The theoretical and methodological assumptions underlying the research were illustrated with a reference to issues concerning linguistic variability and corpus linguistics. Both topics are still to be considered as fundamental for those who, as teachers or learners, deal with English as a foreign language today.

As the study progressed, significant elements in view of further analyses emerged and prompted theoretical reflections on the causal relationship between norm, variability, and error in interlanguage and in second language acquisition in general. The research, therefore, disclosed new perspectives for linguistic analyses which, far from being exhausted in the chapters of the early volume, have led over time to further reflections involving other aspects of language structuring in varieties of English and related methodologies. The present book encompasses indeed the latest developments in that direction.

Linguistic theories are today largely based on ‘objective’ methodologies of analysis which allow researchers to confirm or disprove their initial hypotheses or, vice versa, to draw conclusions from the observation of existing data. The most recent developments in corpus linguistics enhance the latter type of approach (*corpus-driven*), in the belief that *constants* and *variants* of a language should not be observed in abstract situations but rather in relation to authentic communicative acts/contexts. From this perspective, corpus linguistics unfolds a promising scenario for language research, theorisation, and learning.

PART I:

THEORETICAL PERSPECTIVES

Chapter One

Aspects of Linguistic Variability

Introduction

> "The most acceptable sentences are those that are more likely to be produced, more easily understood, less clumsy and in some sense more natural. The unacceptable sentences one would tend to avoid [...].
>
> Acceptability is a concept that belongs to the study of performance, whereas grammaticalness belongs to the study of competence" (Chomsky 1965:11).

The issue concerning the concept of grammaticality/non-grammaticality as well as acceptability/unacceptability of utterances in a given language has long been debated in academic contexts. It involves multiple fields of investigation (linguistics, sociolinguistics, psycholinguistics, pragmalinguistics, cognitive linguistics, glottodidactics), also in consideration of the fact that judgements on grammatical acceptability and appropriateness concern both native and foreign languages. Therefore, the issue will be addressed here by referring to both.

Before proceeding through some of those aspects, it is foremost important to clarify two more issues: the methodological distinction between *competence* and *performance*, and the terminological question concerning the use of the terms *grammaticality* and *acceptability*. The competence component concerns those aspects of linguistic output that are directly conditioned by linguistic knowledge, specifically grammatical knowledge. Performance, instead, includes the production and interpretation of linguistic expressions. The distinction becomes nonetheless problematic when it is applied to purely linguistic phenomena, where, for example, local ambiguity and processing load are seen as factors that may cause difficulties in comprehension (Crocker and Keller 2006).

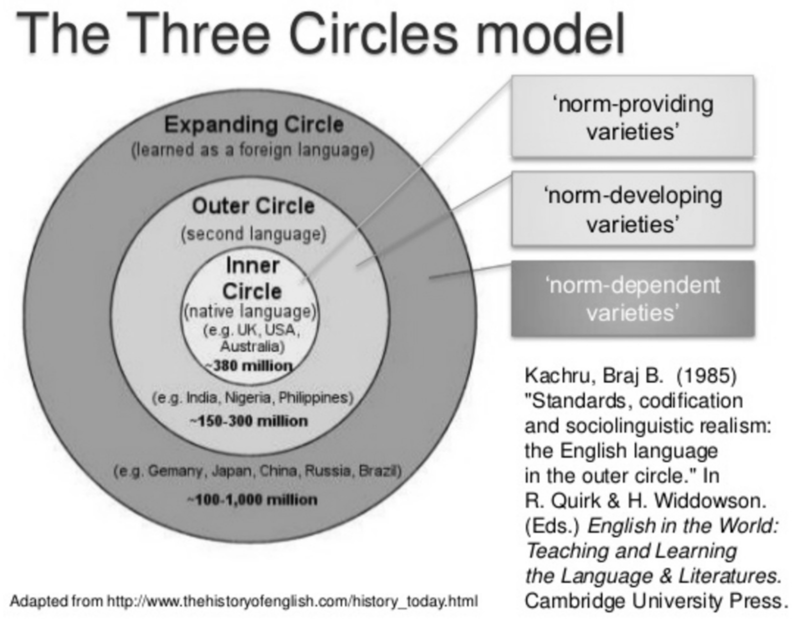
The term grammaticality (in the narrow sense of syntactic grammaticality) is used to refer to the theoretical competence that underlies the performance phenomenon of speakers' acceptability judgements. Acceptability is measured in experiments when subjects are asked to rate the well-formedness of sentences or the probability of their occurrence, which is, in part, determined by factors like sentence length and lexical frequency. It is anyway worth noticing that "grammatical competence is a theoretical entity, which is not directly accessible to observation or measurement. The primary evidence available for ascertaining its properties are speakers' acceptability judgements" (Jey Han Lau et al. 2017, 1204).

Judgments of acceptability and/or grammatical adequacy of utterances are therefore gradient in nature and cannot be directly adapted to a binary formal grammar. This feature represents a crucial topic of linguistic investigation, especially if its aim is the formulation of theories and models of general validity. In this regard, starting from Chomsky's linguistic theories, a central role has been assigned to the native speaker and his linguistic intuitions in theoretical formulations, since he was considered to have access to a quantity of data that could provide indisputable judgements on the formal correctness of model utterances.

The issue becomes even more complex when the utterances of non-native speakers are the very object of analysis, and in particular if we consider the case of the English language that records a continuous increase in the community of those who use English as a Second Language (ESL)[1] or as a Foreign Language (EFL)[2], while the number of speakers of English as a Native Language (ENL or L1)[3] remains fairly stable. The concept of non-native speaker is also currently complicated by the emergence of *New Englishes*[4] in the former British colonies, which tend to create new norms with the consequent *nativization* of English, and the *Englishization* of other world languages' (Kachru 1995, 11f). At the end of the 1980s, the British scholar Braj Kachru provided an extremely effective image to represent the complexity of the status of English in the contemporary age evoking the

[1] "English in countries where it holds special status as a medium of communication. The term has also been applied to the English of immigrants and other foreigners who live within a country where English is the first language" (Crystal 1995:108).
[2] "English seen in the context of countries where it is not the mother tongue and has no special status, such as in Japan, France, Egypt, and Brazil. Well over the half the countries of the world fall into this category" (Crystal 1995:108).
[3] L2 here is used to refer to both ESL and EFL, i.e. "a language which is used in order to meet a communicative need" (Crystal 1995:108), although it is mainly referred to learning English as a foreign language.
[4] "The term is really applicable only when there has been considerable linguistic development away from the traditional standards of British and American English, with some degree of local standardization" (Crystal 2003:313).

shape of three concentric circles to explain its types of spread, patterns of acquisition and functional domains across cultures and languages (ib.):



The Three Circles model

Expanding Circle
(learned as a foreign language)
Outer Circle
(second language)
Inner Circle
(native language)
(e.g. UK, USA, Australia)
~380 million
(e.g. India, Nigeria, Philippines)
~150-300 million
(e.g. Gemany, Japan, China, Russia, Brazil)
~100-1,000 million
'norm-providing varieties'
'norm-developing varieties'
'norm-dependent varieties'

Kachru, Braj B. (1985) "Standards, codification and sociolinguistic realism: the English language in the outer circle." In R. Quirk & H. Widdowson. (Eds.) *English in the World: Teaching and Learning the Language & Literatures.* Cambridge University Press.

Adapted from http://www.thehistoryofenglish.com/history_today.html

Thus, two types of non-native speakers can be readily identified: the former understands and uses English in specific institutional and educational contexts; the latter is represented by the community of speakers of English as a second language who create new norms, thus developing their own institutionalized variety of English (Graddol 2002:67). The emergence of New Englishes has thus led to the codification of new norms spreading across grammars and dictionaries (ib:68) and at the same time to the reconsideration of the theoretical principles of variability/ variation and standard language in an entirely new perspective. More recently, the advancement of variationist studies has led to systematization of theories culminating in Schneider's Dynamic Model (2003; 2007; 2011) in which accomodation processes in language contact situations play a prominent role in developing idigenous forms and Leitner's Habitat Model (2011) which "is able to explain the dual structure of English – the local epicentres (and those varieties that can, in principle, become epicentres) and the global level of international English" (p.33).

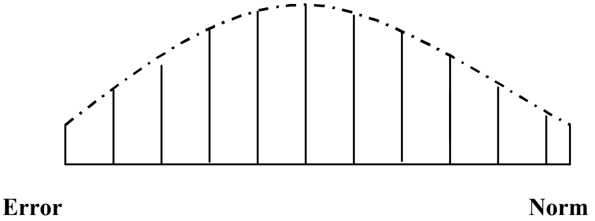
It is in the light of these elements that the characteristics of L2, interlanguage and deviation phenomena from the norm will be observed.

From norm to error

> "What is the nature and what are the parameters of a standard language in a linguistic community? What is the relation between such a standard form and other varieties of the same language? What determines our recognition of a variety in its own right as against a set of variance features that merely indicate imperfect acquisition?" (Quirk 1995, p.vii).

Anyone concerned with interlanguage and error analysis cannot fail to address the same questions with which the British scholar opens the introduction to his book: they highlight aspects that were widely discussed in the past among sociolinguists, but that can be now applied to new problems in the study of the English language. Such issues have been raised by the different status English has been achieving for the last decades within the ongoing process of globalisation.

The following diagram shows the most significant aspects of the phenomena that will be observed in the following sections:



The areas of the graph with different heights represent different degrees between the two ends that are particularly meaningful to researchers in view of a ‘typological’ description of interlanguage. It will be difficult to see how the existence of a number of cases in which speakers’ judgements are robustly binary entails the categorical nature of grammaticality, even when these cases exhibit clearly identifiable syntactic errors that are well described by a given syntactic theory. Gradient judgments will inevitably appear to be sharp for clearly paradigmatic cases.

From this perspective, both the concepts of error and norm are to be seen on a cline that overcomes the apparent opposition of the two ends: they are not entirely opposed terms, but rather expressions of different degrees of grammaticality/acceptability of the same language. Since the terms play an essential role in further discussion about a reconsideration of the phenomenon of variability within the same code, their most salient aspects will be briefly outlined in the following paragraph.

Standard language and standardisation

In linguistics, the expression 'Standard English' refers to the variety most widely accepted and understood within an English-speaking country or throughout the English-speaking world. Its main feature is that "it is more or less free of regional, class, and other shibboleths, although the issue of a 'standard accent' often causes trouble and tension" (McArthur 1996:902). If we compare this definition with the following one, it will be clear that the notion of Standard English is not so easy to define:

> In sociolinguistics, [it is] a much debated term for the variety of English used as a communicative norm throughout the English-speaking world. The notion has become increasingly difficult to handle because of the emergence of different national standards of usage (in vocabulary, grammar, pronunciation and spelling) in areas where large numbers of people speak English as a first or second language: there are important regional differences between UK, the USA, Canada, Australia, South Africa, the West Indies, India, West Africa and several other parts of the English-speaking world (Crystal 2003:431).

The very concept of a standard language is received with suspicion by many (p.7), so that the notion of a standard language is rejected on well-founded historical and linguistic grounds: "There isn't a single all-purpose standard for language" (Quirk 1995, 7s). On this basis, the Standard English of an English-speaking country can be defined as a minority variety (identified chiefly by its vocabulary, grammar and orthography) which carries more prestige and is mostly understood (Crystal 2019, 118).

Standard British English (SBE), widespread within the English-speaking countries and even more extensively within the international community as the most favourite language for its communicative exchanges, has always been identified with its written form, which is highly codified and spread almost uniformly throughout the world. SBE actually comprises three standard varieties: Standard Scottish English, Standard Irish English and Standard English English (used in England and Wales). And it is the latter variety, Standard English English, that is taught as an English language variety (English as a Foreign Language) in formal education contexts (Hughes & Trudgill 1979, 8). Finally, it should be noticed that SBE mainly refers to grammar and vocabulary and includes both formal and informal styles such as *I haven't got any* (formal) and *I haven't got a bloody clue* (informal) (Trudgill 1982:1).

A distinctive feature of standard languages is that they exist at the highest level of abstraction in standardised forms corresponding to completely invariable states of a language. However, only seemingly paradoxical, all languages show a strong tendency towards variability that places them in a state of systematic change. This is equivalent to saying that when one observes a community of standard language speakers, the language they use will never fully conform to it that is just an idealisation of the language.[5] Therefore, it should be noted that when we use the expression 'Standard English' we are not referring to the English language as a whole, because this in turn represents a much more variable and unstable phenomenon (Milroy 1999:18). Thus, dealing with a standard language implies to assume as a superordinate entity, the standard variety used in different communicative situations that tend to support this consideration of the language by speakers.

The most immediate effect of this phenomenon is to make the standard language coincide with the 'correct' use of that language. The notion of 'correctness' thus takes a decisive role in maintaining the ideology of standard language through the formulation of prescriptive rules. It follows that grammaticality/non-grammaticality judgements on a given utterance/sentence are based on its conformity to the rules and constraints of grammars constructed on the basis of specific theories of linguistic description (McArthur 1996:414).

This criterion for correctness identification became central around the 1960s with generative grammar aiming to identify the fixed set of rules that allows the native speaker to distinguish between grammatical and well-formed utterances and ungrammatical utterances deviating from the norm or are poorly constructed. In this context, the concept of acceptability, based on native speakers' judgements about possible/impossible constructions in a given language, took on a precise function.

Closely related to the concept of acceptability is finally that of *variability*, which can be considered not only from a syntagmatic point of view, i.e. on the basis of the possible linguistic realisations in specific standard/non-standard varieties, but also from a paradigmatic point of view considering

[5] "Although SE is widely understood, it is not widely produced" (Crystal 2019, p.118).

the linguistic system itself, i.e. the parameters proper to a given language. Regarding the first point, for example, McArthur (1996:415) points out that according to this principle it may happen that an utterance is grammatical within one regional or social variety and ungrammatical in another.

Allen and Corder (1978) specified that within the same code "differences are not matters of kind but of degree" (1978:67), determined by not only linguistic but also socio-cultural factors. The analysis of the influence of contextual factors on linguistic variation, however, involves some methodological problems. Descriptions of language systems have been based primarily on linguists' assumptions regarding the distinction between 'grammatical' and 'un-grammatical' forms of a language, which are inadequate for a study of language use. Indeed, the study of language use requires corpus-based methodological tools and procedures on a large database of different texts from a statistically significant number of speakers rather than considering a range of contextual factors that determine linguistic variability (Reppen et al. 2002: vii).

Crystal and Davy (1978) also referred to the risk of representing speech acts and characteristics of a given language through a 'one language-one situation' relationship: They considered rather more meaningful "to talk of range of appropriateness and acceptability of various uses of language to given situations" (1978:71). This is the reason why they introduced a *z* factor of unpredictability such as, for example, the occurrence of traits of informality in situations where one would have expected a higher degree of formality. In their scale of language usage, the two scholars represented the formal characteristics of English by placing at one end of the scale the uses of the language in which all possible linguistic forms might occur, and at the other end the uses in which only a limited number of linguistic structures might occur (ib.:73) as determined by specific textual types, such as weather forecasts and press releases. Of course, most grammatical and lexical patterns are imposed on the speaker as rules shared by the whole community of speakers in any communicative situations. Thus, it may happen that some registers, such as 'formal/informal', or text types such as 'religious' or 'legal' texts may share a larger number of linguistic correlates, for example, with informal and colloquial registers, so that for the distinctive linguistic features of a given situation one may speak of stylistic characteristics (or *variety markers*) (ib. :90).

Given the particular attention paid by sociolinguists to the use of language in real contexts, it is now universally acknowledged that languages show forms of 'ordered heterogeneity', linked to the range of registers ap which are propriate to specific uses of the language (Görlach 1999:1). From a diachronic perspective, it is worth noting how variability, seen as a characteristic feature of any language in use, with its regional and social differences, can lead first to the co-existence of variants also indicating different language functions and then to the prevalence of only one of the two allomorphs which will then be taken as the standard form.[6] *Diachronic variation* can be therefore examined in terms of *change* from a long-term perspective.

Socio-cultural models related to *synchronic variation* such as those outlined above might be less relevant in educational contexts of foreign (but not second) language learning since interactional processes only take place in the classroom and among speakers of the same L1.

Finally, we can observe that the concept of variation includes different levels of linguistic analysis: from the occurrence of one grammatical form rather than another (such as the fluctuation in the use of the genitive or nominalisation) to the use of a specific lexical item rather than another, up to linguistic choices determined by style or genre. The combination of these variables may lead to the identification of a linguistic variety proper.

The question arises, however, as to whether the gradual increase in divergence from the source language may lead, as already happened in the past, to languages that are so different that they are no longer mutually intelligible. In contemporary societies, however, such a situation is implausible, since the demands of new information technologies require the sharing of the same code and mutual intelligibility (Quirk 1995:5). Within the European Union, where English is mostly learned as a foreign language, a variety of English, Euro-English, has in fact become widespread, which shows no obvious signs of differentiation that could lead to the formation of mutually unintelligible varieties (Crystal 1997:136). But it is nevertheless possible to assume that the English currently adopted as a lingua franca in the European Union may have characteristics of its own (McArthur 2002:257). As clearly demonstrated by Jenkin's pioneering studies (2000; 2009) the term 'lingua franca' refers to a contact language used among

[6] The codification of one of the variants also demonstrates that nothing new was created in the process of standardisation other than an existing variation that had become grammaticalized according to specific functions (Görlach 1999:9). According to Mattheier (1988), the task of historical sociolinguistics is precisely the study of the selection of variants motivated by the socio-communicative context or by choices of linguistic policy. See also Romaine (1988:1433).

people who do not share a first language and as such it ‘exists in its own right and is being described in its own terms rather than by comparison with ENL’ (Jenkins 2009, 2). Consequently, even errors in an English as Lingua Franca (ELF) framework are not determined by reference to ENL norms, but they should be rather reinterpreted in the light of the role that accomodation plays in communication among non-native speakers of English: any attempt to accomodate which results in a non-native speaker’s form, no matter how intelligible it is to the listener concerned, continues to be regarded as an ‘error’ (ib.: 21).

The main argument here is that as far as ELF is concerned, the so-called ‘errors’ should be considered legitimate fetures of NNS’s English which may respond to demands of projecting L1 identities in L2 English (ib.: 23).

Finally, it should be noted that different taxonomic principles are often adopted when analyzing varieties, so that it becomes necessary to specify a precise taxonomy that identifies varieties starting from the concept of *use* and *user*. Then, it will be possible to draw a further distinction between native varieties (such as the institutionalized varieties of British and American English) and non-native varieties of English leading to the identification of ‘Russian/French/German English’ and so on, also defined as ‘performance varieties’, whose main problem lies in their fundamental instability (Quirk 1995, 24).

Language between encoding and variation: pattern grammar and language creativity

The phenomenon of language variation and variability has been so far considered as the result of the interaction of different extra-linguistic factors.

But the phenomenon can also be analysed from a more general and more strictly linguistic and theoretical point of view, by comparing theoretical models, apparently distant from each other but, arguably, internally correlated.

One of the most recent explanatory models, the so-called *pattern grammar* (Francis & Hunston 2002), has been proposed by the authors as a completely new approach to linguistic description, aiming at reinterpreting some aspects of traditional grammar but focusing on the linguistic bearing of individual lexical items.

Pattern is defined as a sequence of functional words, lexemes (word types) or clause types frequently co-occurring with a given lexical item. These frequency patterns can be observed through lists of concordances created by specific software (concordancers) (see later ch. 2). These concordance lists give researchers the considerable advantage, among other things, of making even more evident the close link between lexicon and grammar as well as the eminently phraseological, rather than lexemic, nature of the rules of language choices and uses.

The theorists of this model, therefore, insist on a linguistic description that privileges the analysis of single lexical items, thus distancing themselves from non-corpus-based linguistic theories. The importance given to the concept of pattern in this model constitutes an interesting area of investigation that unifies the domanis of lexicon, grammar and meaning (Hunston 2002, 166). In fact, the identification of some recurrent patterns in a given variety or register can help, for example, identify their original meanings. Since the same meaning can often be expressed through a variety of combinations of the *lexicon-pattern* system, it becomes necessary to know how variation across those combinations may correlate with certain variations in register (ib.). For example, some studies have examined specific verb complementation patterns which may require a different distribution of finite and non-finite clauses such as *that*-clauses (particularly frequent in spoken discourse), *to*-clauses or *ing*-clauses (more frequent in written discourse) depending on the register used (Biber et al. 1999). Thus, the main assumption of this theory is represented by the selection properties of individual lexical items and is actually already considered within the traditional syntactic theory: its innovative element is rather represented by the corpus-based methodology adopted in linguistic analyses.

It should be noted, however, that earlier in 1965 Chomsky in his *Aspects of the Theory of Syntax* (p.101) had highlighted the systemically closed nature of language, which can be exemplified by constructions such as *he decided on the boat*. This simple sentence can be interpreted either as 1. *he made the decision to buy a boat* or 2. *he made a decision while on the boat.*[7] 'Closed' constructions referring to the first interpretation are examples of *grammatical*

[7] In the former interpretation, the prepositional phrase is linked to the main verb, whereas in the latter, the phrase containing the adverb of place can occur in a free position with respect to the verb.

collocations, while those belonging to the second interpretation are examples of *lexical collocations.*[8]

A grammatical collocation is a recurrent combination consisting of a dominant or head word such as verb, noun or adjective followed by a grammatical or functional element such as propositions like V+Prep (e.g. *aim at*), N+Prep (e.g. *interest in*) or ADJ+Prep (e.g. *afraid of*). A lexical collocation does not have subordinate elements because lexical elements of equal relevance such as ADJ +N, N+V, V+N and Prep+N may co-occur. Many of these combinations formed by V+N represent ‘open’ combinations, in the sense that, for example, the verb *to build* can be followed by nouns such as *house, bridge* or *road* without restrictions on choice by speakers. Collocations of polysemic words are extremely useful for disambiguating their meaning, as in the case of the word *line*, which can be collocated with either the verb *draw*, meaning ‘to draw a line’, or *form*, with the meaning of ‘to align’.

Collocations therefore represent a fundamental element of lexical compositionality as they determine the acceptability judgements of syntactic constructions. They should not be confused with idiomatic constructions, which are also fixed expressions, but structured in such a way that the meaning of individual elements composing them has no relationship with the meaning of the whole construction: see, for example, the expressions *to have one’s back to the wall* or *to kill two birds with one stone.*[9] Proverbs (e.g. *an apple a day keeps the doctor away*) differ from the former typology, as they consist of complete constructions that, unlike idiomatic expressions, enable less lexical and syntactic variability.

Therefore, lexicalised constructions occurring in any natural language tend to have a very regular thematic structure. Modifying processes can naturally intervene on those codified structures and make the constructions formed by those encoded by the lexicon be enriched with new expressions. If we wanted to extend the discussion to include the semantic aspects that such an analysis involves, then we could not avoid considering the concept of *head valency* and the number of *arguments* with which it can be combined to

[8] For further discussion of the concept of ‘collocation’ and its central role in corpus linguistics, see Chapter 2.

[9] The first expression with the meaning ‘to be in a very difficult situation’ can of course be modified by using, besides different subjects, also a transitive verb such as ‘to get sb. into a corner’. In the second expression, meaning ‘to achieve two things by doing a single action’, variability only concerns the choice of subject.

form a complete sentence. The fact that each lexical item has a certain valency allows us to explain the non-grammaticality of some utterances with respect to others and also the particular type of relationship between a lexical item, be it a verb, a noun, an adjective or a preposition and its arguments, called *semantic* or *thematic roles*.[10]

These issues are of considerable relevance for the debate on the underlying mechanisms of (both L1 and L2) language acquisition, as they lead us to reflect on the mechanisms by which speakers perceive the phraseological structures and individual lexical items of L1 and L2.

In fact, from a native speaker's point of view, language appears as a set of recurring structures organised in certain fixed or semi-fixed strings, in which words show a reciprocal relationship, linked to each other in a more or less variable way, without, however, seemingly constituting a single lexical entity. This approach to language interpretation and decoding, which we might call 'phraseological' rather than 'lexematic', refers to theorethical models, such as those proposed by Sinclair (1991), which provide a different perspective for understanding the syntactic structures of a language: the 'idiomatic' principle, similar to the one proposed earlier, and the 'open choice' principle, with the result that the meaning of an utterance is derived either from the sentence as a whole according to the most conventional and recurrent phraseology or from the meaning of individual lexemes that operate according to well-defined grammatical rules (Hunston 2002, 145).

In this perspective, the *lexical priming hypothesis* proposed by Hoey[11] appears to be highly suggestive, because it goes beyond the concept of collocation as "the statistical tendency of words to co-occur" (Hunston 2002, 12). According to this theory, every single lexical item is selected for its use in discourse in a given grammatical role as a result of the communicative effect of the speaker's encounter with that particular item. Since this lexical 'selection' (priming) in natural languages shows some drifts, its scope of interest must also include the sphere of knowledge, gender and addressee, leading to a distinction between 'receptive priming'

[10] The thematic role in Chomskyan terminology is referred to as theta-role (θ-role), while 'theta-criterion' expresses the restrictions imposed by lexical properties on syntactic representations (Chomsky 1986).

[11] The following lines summarises some of the arguments presented by the linguist himself during a plenary session of the 24th ICAME Conference held in Guernsey in April 2003.

and ‘productive priming’. Grammatical constructions will, therefore, reflect the selection of the most common words and lexical combinations that are considered as the speaker’s ability to recognise certain patterns in a given language.[12] At the same time, however, features of originality play a relevant role and introduce an element of language variation and creativity[13] that is realised through the modification and/or fusion of certain groups of words (*chunks*) to satisfy communicative needs adapted to specific contexts without, however, creating problems of production and comprehension that could be generated by the new linguistic combinations (Barlow 2000).

From an L2 point of view, the analysis of learners’ phraseology is extremely interesting, as it contributes to a better understanding of cognitive and processing mechanisms operating in any individual’s mind. In fact, if language is a set of structures corresponding to specific meanings,[14] it would be necessary to question why this principle is not immediately transparent even to non-native speakers and explain the phases through which the representation of meaning for certain constructions is construed. One possible explanation for this mechanism of language ‘processability’ and meaning construction in non-native speakers lies in their general tendency to infer meanings from individual words and not from larger syntactic constructions. In this regard, Corder (1981) had already found a tendency in foreign language learners to implement the same strategy as in children, i.e., to search for meaning through the analysis of the most relevant elements, such as single items or lexical strings.[15]

An interesting model for explaining the processes underlying language learning is the cognitive model proposed by Ellis (1996), based on the learner’s ability to select and memorise language patterns. This model takes into account both the learner’s creative knowledge of linguistic rules and the lexical strings stored in the learner’s mind. It follows that the acquisition

[12] This hypothesis echoes George Zipf’s ‘Minimum Effort Principle’ (Manning et al. 2003:27) according to which “conservation of speaker effort would prefer there to be only one word with all meanings, while conservation of hearer effort would prefer each meaning to be expressed by a different word”.

[13] This principle seems to refer to the principle of *recursivity* theorised by generative grammar, according to which the rules of recursivity represent the main formal means of accounting for the creativity of language: through this mechanism it is possible to produce an infinite set of sentences from a finite set of rules.

[14] According to Sinclair, language consists of pattern and meaning in the sense that what is traditionally called ‘grammar’ can also be called ‘pattern’.

[15] According to Corder, a lexical string does not necessarily have to be understood as a lexical group or chunk with a definite meaning.

of both a first and a second language is essentially the acquisition and analysis of memorized lexical strings following repeated exposure to the target language. Thus, the learner acquires a basic vocabulary in which words can be seen in relation to others with which they co-occur regularly. In the case of a first language, an automatic and implicit process of analysis of these lexical combinations is activated, which then results in an abstract formulation of grammatical regularities. L2 learners, on the other hand, perceive an almost completely explicit knowledge of the target language grammar which, along with the limited lexical repertoire at their disposal, is reflected in a wider use of grammatical rules or in a 'misuse' of memorized strings in their language performance.

It is evident, therefore, that the nature of the lexicon / grammar relation represents a crucial node in this context. In order to analyse this aspect, it is useful, once again, to refer back to 1960s linguistic theories and to Lyons (1969:171) in particular:

> Many older books on language devoted a good deal of space discussing which of the two traditional primary units of grammatical description, the word or the sentence, is to be regarded as 'basic': does the grammarian first of all identify words and then account for the structure of sentences in terms of the permissible combinations of words, or does he start by recognising the sentences in his material and then analyse the sentences into their constituent words?

Later on, he will explain that the sentence represents for the linguist the largest unit of analysis able to explain the distributional relations between the elements of an utterance. In Saussurian terms, utterances are strings of words generated by a system of lexical elements and rules that constitute the *langue*, through which the linguist describes examples of *parole*. As for the distinction between grammatical and lexical collocations on the one hand and idiomatic and proverbial expressions on the other, already mentioned above, Lyons also proposes to consider them 'incomplete sentences' because they do not allow any extension or variation. But he also points out that they do not respond to the rules of completeness prescribed by grammar, since they are learned as unanalysable units or sets, employed on certain occasions by native speakers (p.177).[16] Along with

these types, Lyons also considers constructions that are unstructured or only partially structured, but which can nevertheless be combined into sentences according to the productive rules of the grammar of a language: these are

[16] An example of this type proposed by Lyons is *all that glisters is not gold.*

what he calls *schemata* like *What's the use of -ing*? from which a large number of utterances can be generated by filling the empty space in the schema with an element of the appropriate grammatical class.

Comparing the different general theories that have followed in recent years and that have been based on computerised analyses of linguistic data, two elements can be identified that represent both a point of contact and differentiation among theories. The first element concerns terminology: where sentence-schemata and phrase-schemata were used to define the types of utterances described above, the terms frames / fixed expressions / semi-fixed expressions were later adopted (Renouf et al. 1991; Lewis 1993). The second feature is the apparent 'detachment' with respect to the past: analysing the range of the most recent theories based on the analysis of linguistic data with the aid of information technology, it is surprising that in most of the theoretical reflections on these aspects of languages there is neither specific nor indirect reference to previous linguistic theories, despite the fact that they have touched upon the same issues and often formulated hypotheses that the data of more recent research have not disproved so far.

Variability in interlanguage and its pedagogical implications

Variability in interlanguage[17] is even greater, both because of the nature of the L2 learning process and possible mother tongue interference. In fact, it may happen that the L2 learner in certain contexts switches codes at the level of proposition, sentence or discourse (Perlman Lorch 1995:255). The mechanism of code-switching, that is, the ability to switch from one language to another, is to be considered, from a cognitive point of view, similar to the native speaker's ability to switch from one register to another (ib:257).

What is clear, however, is that learners manifest considerably different levels of variability in their L2 productions, leading to questioning whether these phenomena are simply performance factors or whether they are rather

[17] The term interlanguage refers to "the linguistic system created by someone in the course of learning a foreign language, different from either the speaker's first language or the target language being acquired. It reflects the learner's evolving system of rules, and results from a variety of processes, including the influence of the first language ('transfer'), contrastive interference from the target language, and the overgeneralization of newly encountered rules" (Crystal 2003:239).

the effects of an underlying system in formation. It is likely that learners construct variable systems in an attempt to identify particular forms that perform specific functions (Ellis 1998:28). Of course, such form-function patterns produced by learners do not always conform to those required by the target language, since they do show a certain systematic occurrence that points to the hypothesis of the existence of a variable system of form-function mappings.

Several hypotheses have been put forward in this regard: some scholars suppose that interlanguage variation is free because, within a range of possible choices available to the L2 learner, he will alternatively use two or more forms for the same function or, at an early stage of L2 acquisition, the same form may be used to express different functions.

This issue has, predictably, important implications for the teaching of English as a foreign language.

In past Western tradition, the written texts of classical Greek or Latin literature constituted the norm and the model to be followed in order to successfully achieve pedagogical objectives, so much so that the norm to be taught was considered to correspond to a ‘prescribed’ variety of the language.

Therefore, standardised forms of the written language as they appeared in the canonical works of literature were also assumed as the norm in foreign language teaching (Kramsch 2002:59). Until the 1950s, the teaching of grammatical structures, often presented by means of contrastive analysis between L1 and a target language, was still widespread, with the clear aim of developing a new linguistic attitude that was as error-free as possible. In this way, little attention was paid to the standard language actually used in real communicative contexts, whereas factors such as the type of relationship existing among speakers, their social status and the context of interaction in which each communicative act takes place were relevant, as had been highlighted by Labov’s research on language variation.

Even in the 1970s, research on foreign language acquisition did not consider the native speaker as a model, but rather the national standard language represented by its grammar, paradigms and written texts. Thus, learners’ language skills were assessed according to the grammar and spelling rules of the nationally recognised language code (Kramsch 2002:61).

Gradually, also in the field of language teaching, language variability began to be considered in the design of teaching materials and activities, together with the notion of native speaker norms, which led to the creation of authentic tasks based on actual language use and pragmatic appropriateness.

In emphasising the authenticity characteristics of native speaker's norms, research on foreign language learning has highlighted the rich social context associated with language use and has consequently analysed the learners' acquisition process by also considering the verbal interaction mechanisms at work in everyday L1 contexts (Kramsch 2002:60). It has been observed that the typology of communication usually adopted by foreign language learners is pragmatic rather than syntactic: in the first type, canonical theme-syllable structures are preferred, little use of conjunctions that increase the incidence of nouns over verbs, syntactic order governed by the communicative context; in the second type, SUBJ+V structures, subordinate constructions, greater use of connectors and verbs are predominate (Kerr 2002:186).

Pragmatic constructions as well as clefts[18] appear more frequently in spontaneous speech, in which the 'new' element of the utterance appears in the first part of the clause and the 'given' element in the second one to simplify the mechanisms of the message processability.

In fact, these structures seem to appear first as models in learner's interlanguage, in which the syntactic constituents are juxtaposed to each other in a linear and ordered sequence of theme-rheme constructions. This pattern increases lexical density to the detriment of a greater cohesion of the utterance. Other constructions containing typical features of the spoken language may also appear in the interlanguage, but are not acceptable in the written language, such as contracted forms ('*cause*), independent propositions joined by coordinating conjunctions that reproduce a repetitive syntactic structure (*And I...And she...*), use of the present tense to indicate past actions (Gass 2002:242).

In recent years, the interest of researchers in the field of language teaching has been shifting more and more towards the acquisition and enhancement of learners' metalinguistic skills, considered as important as communicative competence in L2. Metalinguistic competence implies, among other subskills,

[18] Examples of those structures are utterances such as «and there's a man there, he had, there was a man there, who was playing, the boy is looking at the fish that, that were the prize for the game» which are typical of spontaneous speech (Kerr 2002:187).

the ability to reflect on the potential variability of language and the different possible relations between form and meaning.[19]

Therefore, if one accepts variability as an element to be included within the set of pedagogical norms of the target language to which learners are exposed, then it will necessary to recognise that different contexts require variable norms: learners in academic contexts will learn, therefore, linguistic norms that are presumably different from those of learners who will use the L2 in different contexts, such as commercial transactions. The only apparently antithetical nature of the expression ‘variable norm’ is explained by the demand for variability and change, assumed as essential elements of teaching practice increasingly linked to changes in language, which are determined by the diversification and extension of the contexts of use generated by the ongoing process of globalisation.

The relevance of the change brought about by the new concept of linguistic norm cannot but have immediate effects also on the concept of ‘error’ that needs, consequently, some revision and appropriate clarification. It is in fact necessary to refer also to the concept of variable norm when analysing learners’ errors. They should actually be classified not only on the basis of grammatical categories, but above all on usage constraints (determined by written or spoken registers), the linguistic materials the learner has been exposed to and the criteria of adequacy to the proposed task upon which the acquired linguistic competences are evaluated.

The concept of error in foreign language learning research

In the light of the arguments illustrated so far, it is clear that the concept of error needs an in-depth examination, especially for some relevant aspects that have been unjustifiably neglected until now in research concerning error analysis:

[19] The analysis of variation in interlanguage proposed by Dewaele is particularly interesting to explain specific phenomena occurring in spoken interlanguage (1995:235) introducing the concept of explicitness/implicitness related to deixis. The main characteristic of formal language would consist in a higher level of explicitness due to the need not to create ambiguity in the message that would require the addition of information. In the frequent use of spatio-temporal references directing the interlocutor’s attention to precise contexts, the need for the learner to add further information becomes evident. On the contrary, the less formal style uses more frequent lemmas which are therefore acquired more easily and quickly.

1. the linguistic materials learners have been exposed to during classroom teaching activities; this can be distinguished by varieties of style, register, lexical/grammatical density, i.e. factors that provide important information on the learner's preferences and appropriateness of use of certain structures;

2. the language used as a model for the analysis of the interlanguage; one can in fact assess the conformity to the rules of prescriptive grammar or take the variety of the spoken standard language as a reference point;

3. the relevance of quantitative and qualitative data, which can help to better define the distinction between 'error' and 'mistake' and the role of avoidance mechanisms as important feedback indicators.

In particular, for the purpose of a comprehensive survey, it is crucial to have the appropriate tools to distinguish between mistake and error. To this end, it may be useful to monitor the frequency of a certain phenomenon of deviation from the norm over a given period of time or simply to induce the learner to self-correct (Ellis 1998:17). However, the occurrence of phenomena of variability in the interlanguage does not always help provide definitive answers. The identification and description of errors therefore appear to be a crucial issue.

There are different ways of classifying errors: on the basis of grammatical categories by putting together, for example, all the errors related to the category of 'verb' such as number, diathesis, tense; or on the basis of the comparison between learners' linguistic productions and those from native speakers'. However, the assessment of errors also depends on the linguistic competence of the investigator and is therefore exposed to the risk of subjective evaluations of learners' utterances.

The same system of labelling errors according to grammatical categories (lexical-grammatical, morphological, syntax; see next chapter) sometimes seems incomplete because grammatical categories are not discrete classes in either L1 or L2 and therefore it is often difficult to label certain errors.

In any case, the fact that errors occur with a certain systematicity suggests the construction of a system of rules by the learner. Errors often have traits that can be generalised, so that they become characteristic of all learners regardless of their L1. This has led some scholars (Dulay et al. 1985:261-296) to hypothesise an order of acquisition of grammatical structures. It is precisely on the basis of this order that errors such as the omission of certain elements within utterances or the generalised extension of a certain linguistic feature (e.g. the -ed morpheme of the past tense of irregular verbs extended also to irregular verbs as in **goed*) can be explained. Other errors relate to interference mechanisms from the mother tongue on the structures of the L2 being acquired.

All these, however, are indicators of a single phenomenon: the process of reorganisation and reconstruction taking place within the learner's linguistic system(s), which is absolutely predictable and can be described in a scheme that, in proposing this process of acquisition, takes the typical U-shaped: at an initial level learners show a certain degree of formal accuracy in their linguistic production, immediately followed, however, by a period of apparent regression to then approach the norms of the target language. This U-shaped pattern may be accounted for the phenomena of variability in the interlanguage (see previous paragraph) and the alternation of correct and incorrect forms (Ellis 1998:23).

Regardless of the nature or origin of errors, it is evident that their frequency in learners' language points to the construction of new cognitive categories that constitute what can be defined as the learner's emerging 'linguistic awareness' (Murray 2002, 187).

The concept of error in variationist research

If we shift the perspective from the individual (learners') to a more general (speech communities') level, the focus will be on understanding what happens to languages when speakers come into contact for different reasons. Given the huge literature on language contact, it is important to focus on specific issues concerning the status of English as a second language to speakers from different first language backgrounds. For example, contact between spoken and written language would lead to the emergence of new spoken norms and new spoken standards which combine structural and lexical elements of two different linguistic systems (Haugen 1972, p.57). One of the main issues to be addressed will be therefore the clarification of the historical interactions and language contact phenomena occurring between pre-existing language standards and the emerging standard language which makes the process of standardization as a special type of language change within a more general theory of language contact (ib.). A research perspective which pays attention to language feature selection and convergence at different historical times would contribute to a better understanding of the nativization and standardization processes emerging varieties of English have been undergoing during the last century. The variationist perspective adopted is therefore far from considering varieties of English as 'deviant versions' of Standard English and local peculiarities, if any, are not described as deviations or errors with respect to 'the codified norms of EnglishEnglish' (Trudgill & Hannah 2001) which may lead to interpret them as 'deviant learner variety/ies' (Sedlatleck 2009, 31). By interpreting them as varieties of their own, instead, with their proper linguistic development and evolutionary path across centuries may contribute to the understanding of the dynamics of language formation and change in complex multilingual contexts where old (English) and new exonormative languages continuously interact with one another and with indigenous, endonormative languages. In Mufewene's (2001, 2005) theory of language ecology and evolution the emergence of contact varieties can be regarded as a series of selective actions from 'a pool of linguistic variants' from different language backgrounds and experiences available to speakers in a contact setting (Schneider 2007, 22). The 'natural' selection of specific features as stable elements of the new varieties depends on both extralinguistic and intralinguistic factors such as the nature of the linguistic input elements, typological similarities between the languages involved in the process of competition and selection as well as the influence of social determinants (Schneider 2007, 23). A new, even though controversial, theory on the formation of colonial Englishes has been proposed by Trudgill (2004, 1-3) in a deterministic view of language change and gradual variety differentiation. In particular, he supposes the combination of six factors to explain the differences between new/post-colonial Englishes and British English which can be briefly outlined as follows:

1. Adaptation to a new physical environment
2. Linguistic changes in the mother country
3. Linguistic changes in the colony
4. Language contact with indigenous languages
5. Language contact with other European languages
6. Dialect contact determined by the different geographical and social origins of the settlers

In relation to factor 2, Mukherjee and Gries (2009) take the development of Indian English (IndEng) as an example and note that "the input variety of British English is a diachronically changing reference point" (p.34) so that "for any comparison of a New English variety with its historical input variety a diachronic corpus of the native variety is needed" (p.35). As regards factor 4, still referred to as IndEng, English came into contact with Indic and Dravidian languages as a consequence of the colonial expansion from the early seventieth century on and the contact variety which emerged in this setting served as lingua franca to ethnically and linguistically diverse populations (Winford 2003, 242). With reference to factor 5, it is important